BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37147
doi: 10.3897/biss.3.37147

OPEN ACCESS

Conference Abstract

Addressing Uncertainties in Machine Learning Predictions of Conservation Status

Barnaby E Walker‡, Tarciso Leão‡, Steven P Bachman‡, Eve Lucas‡, Eimear Nic Lughadha‡

‡ Royal Botanic Gardens, Kew, London, United Kingdom

Corresponding author: Barnaby E Walker (b.walker@kew.org)

Received: 11 Jun 2019 | Published: 18 Jun 2019

Citation: Walker B, Leão T, Bachman S, Lucas E, Nic Lughadha E (2019) Addressing Uncertainties in Machine Learning Predictions of Conservation Status. Biodiversity Information Science and Standards 3: e37147. https://doi.org/10.3897/biss.3.37147

Abstract

Extinction risk assessments are increasingly important to many stakeholders (Bennun et al. 2017) but there remain large gaps in our knowledge about the status of many species. The IUCN Red List of Threatened Species (IUCN 2019, hereafter Red List) is the most comprehensive assessment of extinction risk. However, it includes assessments of just 7% of all vascular plants, while 18% of all assessed animals lack sufficient data to assign a conservation status. The wide availability of species occurrence information through digitised natural history collections and aggregators such as the Global Biodiversity Information Facility (GBIF), coupled with machine learning methods, provides an opportunity to fill these gaps in our knowledge. Machine learning approaches have already been proposed to guide conservation assessment efforts (Nic Lughadha et al. 2018), assign a conservation status to species with insufficient data for a full assessment (Bland et al. 2014), and predict the number of threatened species across the world (Pelletier et al. 2018).

The wide range in sources of species occurrence records can lead to data quality issues, such as missing, imprecise, or mistaken information. These data quality issues may be compounded in databases that aggregate information from multiple sources: many such records derive from field observations (78% for plant species in GBIF; Meyer et al. 2016) largely unsupported by voucher specimens that would allow confirmation or correction of their identification. Even where voucher specimens do exist, different taxonomic or

geographic information can be held for a single collection event represented by duplicate specimens deposited in different natural history collections. Tools are available to help clean species occurrence data, but these cannot deal with problems like specimen misidentification, which previous work (Nic Lughadha et al. 2019) has shown to have a large impact on preliminary assessments of conservation status.

Machine learning models based on species occurrence records have been reported to predict with high accuracy the conservation status of species. However, given the black-box nature of some of the better machine learning models, it is unclear how well these accuracies apply beyond the data on which the models were trained. Practices for training machine learning models differ between studies, but more interrogation of these models is required if we are to know how much to trust their predictions.

To address these problems, we compare predictions made by a machine learning model when trained on specimen occurrence records that have benefitted from minimal or more thorough cleaning, with those based on records from an expert-curated database. We then explore different techniques to interrogate machine learning models and quantify the uncertainty in their predictions.

Keywords

IUCN Red List, machine learning, natural history collections, uncertainty, conservation assessment

Presenting author

Barnaby E Walker

Presented at

Biodiversity_Next 2019

References

- Bennun L, Regan E, Bird J, van Bochove J, Katariya V, Livingstone S, Mitchell R, Savy C, Starkey M, Temple H, Pilgrim J (2017) The value of the IUCN Red List for business decision-making. Conservation Letters 11 (1): e12353. https://doi.org/10.1111/conl.12353
- Bland L, Collen B, Orme CDL, Bielby J (2014) Predicting the conservation status of data-deficient species. Conservation Biology 29 (1): 250-259. https://doi.org/10.1111/cobi.12372
- IUCN (2019) The IUCN Red List of Threatened Species. Version 2019-1. https://www.iucnredlist.org/

- Meyer C, Weigelt P, Kreft H (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecology Letters 19 (8): 992-1006. https://doi.org/10.1111/ele.12624
- Nic Lughadha E, Walker B, Canteiro C, Chadburn H, Davis A, Hargreaves S, Lucas E, Schuiteman A, Williams E, Bachman S, Baines D, Barker A, Budden A, Carretero J, Clarkson J, Roberts A, Rivers M (2018) The use and misuse of herbarium specimens in evaluating plant extinction risks. Philosophical Transactions of the Royal Society B: Biological Sciences 374 (1763). https://doi.org/10.1098/rstb.2017.0402
- Nic Lughadha E, Graziele Staggemeier V, N. C. Vasconcelos T, Walker B, Canteiro C, Lucas E (2019) Harnessing the potential of integrated systematics for conservation of taxonomically complex, megadiverse plant groups. Conservation Biology https://doi.org/10.1111/cobi.13289
- Pelletier T, Carstens B, Tank D, Sullivan J, Espíndola A (2018) Predicting plant conservation priorities on a global scale. Proceedings of the National Academy of Sciences 115 (51): 13027-13032. https://doi.org/10.1073/pnas.1804098115